

# Supplementary Information

## 1 Optimization algorithm

---

Algorithm 1: Alternating Least Squares

---

```
1: Input:  
    $k$  : number of latent factors  
    $\Gamma$ : pairs of entities for initialization  
   For every task  $t$ ,  
    $\{\mathbf{x}_{ti}\}, \{\mathbf{y}_{tj}\}$ : feature vectors for pathogen and host proteins resp.  
    $\Omega_t$ : the observed entries of the matrix  $M_t$   
2: Initialization:  
3: An iteration  $r$ , let  $\mathbf{S}^r$  represent  $\{S_t^r\}_{t=1}^T$   
4:  $\mathbf{S}^0 \leftarrow 0$   
5:  $U^0 \leftarrow$  top- $k$  left singular vectors and  $V^0 \leftarrow$  top- $k$  right singular vectors from the SVD of  $\sum_{(i,j) \in \Gamma} \mathbf{x}_i \mathbf{y}_j^\top$   
6:  $\mathcal{L}^0$  : initial loss  
7: repeat  
8:    $U^{r+1} \leftarrow \underset{U}{\operatorname{argmin}} \mathcal{L}(U, V^r, \mathbf{S}^r)$   
9:    $V^{r+1} \leftarrow \underset{V}{\operatorname{argmin}} \mathcal{L}(U^{r+1}, V, \mathbf{S}^r)$   
10:  For each task  $t$   
     $S_t^{r+1} \leftarrow \underset{S_t}{\operatorname{argmin}} \mathcal{L}(U^{r+1}, V^{r+1}, \mathbf{S}_{-t}^r)$   
11:  Compute  $\mathcal{L}^{r+1}$  and let  $\delta \leftarrow (\mathcal{L}^r - \mathcal{L}^{r+1})/\mathcal{L}^r$   
12: until  $\delta < \tau$ 
```

---

## 2 Parameter tuning

We tune the hyper parameters using a 3 fold cross-validation (CV) on the training split. For all baseline regularization parameters we tried the range:  $[100, 50, 10, 1, 0.1, 0.01, 0.05, 10^{-3}]$ . To address the class-skew we assign a higher weight to the positives. For BSL-MTL, to tune the rank parameter ' $k$ ' we tried:  $[1, 5, 10, 25, 40, 60, 100]$  and the regularization parameter controlling the norm of  $U$  and  $V$  was tuned over the range  $\lambda = \{1 \dots 10^{-3}\}$ . For each task  $t$ ,  $\sigma_t$  was varied over the values  $[10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}]$  and  $\mu_t$  was varied over  $\{0, 0.25, 0.5, 0.75, 1\}$ . The optimal setting was:  $k = 10$ .

### 2.1 Disadvantages of the Homolog method

In Figure 1 we show for each task, the protein interactions in the test data (for one train-test split) that were not found by the homologs model which relies purely on homology-based sequence similarity. The interactions mainly involve human proteins which do not have a homolog in the training data. This illustrates that our features capture more information than mere sequence similarity. Further, inclusion of interactions from other viruses in the naïve manner as done by the homologs in fact hurts the performance (results in too many false positives) - our model is able to learn what to borrow from the other tasks accurately.

## 3 Biological significance of interactions

We analyze sequence motifs derived from the top  $k$ -mers that contribute to interactions. The significant entries of the model parameters  $U$ ,  $V$  and  $\{S_t\}$  were used to compute these motifs.

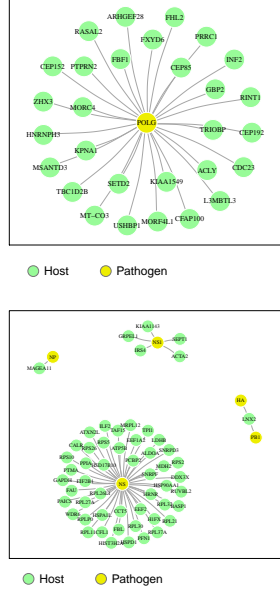


Figure 1: PPI from each task that are found by our method but are missed by the homolog method. To obtain these, we chose a random train/test split and picked examples from the test data on which the homolog method fails - these are ones for which sequence-similar proteins are not found. (a) shows *Ebola*-human PPI, (b) *Hepatitis-C* human PPI and (c) *Influenza*-human PPI.

The top positive-valued entries from the product  $UV^T$  indicate which pairs of features:  $((f_v, f_h):$  virus protein feature, human protein feature) are important for interactions across all the virus-human PPI tasks. Analogously, the entries from  $S_t$  give us pairs of features important to a particular virus-human task ‘ $t$ ’. We find that most of the top entries from  $UV^T$  correspond to linear virus features, whereas those from the various  $S_t$  involve bilinear features.

We analyze the  $k$ -mers corresponding to the top 20 features from each of the matrices.

Note that our features do not directly correspond to a unique amino-acid  $k$ -mer (see Section 4.2): the virus feature  $f_v$  will map to several amino-acid sequences (for instance KKCC, KRCC, RRCC etc all map to a single feature due to the molecular similarity between the amino acids K and R being both positively charged). Given the set of top virus features we can obtain the corresponding set of amino-acid  $k$ -mers, say  $AA_v$ , by reversing the feature-generation step. However most of the possible  $k$ -mers do not appear in the training data (ex: out of the 160,000 ( $=20^4$ ) possible 4-mers  $\approx 24,000$  appear). Let  $AA_{tr}$  be the set of amino-acid  $k$ -mers that appear in the training data. Then, the intersection  $I_v = AA_v \cap AA_{tr}$  gives us the important amino-acid  $k$ -mers from virus proteins w.r.t interaction prediction. To summarize  $I_v$ , we use a popular tool Seq2Logo to generate a sequence motif. The logos for the two-, three-, four-mers from  $I_v$  are generated independently. Since we only want to summarize, we use the Shannon logo type (which does not consider any background amino-acid distribution) with the following settings: clustering-method=None, weight-on-prior=1 (pseudo-counts do not make sense in our analysis). Figure 2 shows the motif that is common across viruses. We observe that the shared tri-mer motif for virus proteins in Figure 2 is dominated by hydrophilic amino acids (T, K, R, D, E).

This procedure described above is used to analyze the most significant human protein features, obtained from the matrix  $UV^T$ . These motifs are shown in the supplementary. The task-specific features i.e significant  $k$ -mers from virus-proteins of *Ebola*, *Hepatitis* and *Influenza* are obtained from the matrices  $S_{ebola}$ ,  $S_{hepc}$  and  $S_{flu}$  respectively. The motif for Hepatitis is shown in Figure 2(right); the rest can be found in the supplementary. The virus-specific motifs seem to be dominated by hydrophobic residues (I, P, L, V, A, G) though S and T do appear in

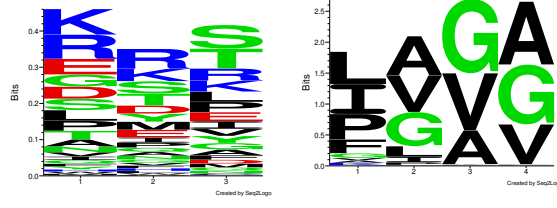


Figure 2: (Left) Top tri-mer sequence motifs from virus proteins across all three viruses. These correspond to features important for interactions and is derived from our model parameter  $H = UV^T$ . (Right) Motif corresponding to enriched 4-mer patterns found by our model (parameter:  $S_{hepc}$ ) for the Hepatitis-C virus task.

some motifs as well.

*Phosphorylation sites:* We found the frequent occurrence of S and T and sometimes Y in the motifs striking and suspected this may be related to the amino acids being frequent targets of phosphorylation. Phosphorylated sites often serve as PPI sites, and databases such as Phospho-site are repositories for known sites in human proteins. Since these are sites in human proteins, we searched for the patterns from the 4-mer motif in Fig 6 and found several to be flanking known phosphorylation sites in human proteins: LLLs, LLLt, ILLs, PPPs, PIPs, PIPt, LIPs, PLLt (lower-case indicates the putative phosphorylation site). This observation also supports the notion that the motifs predictive of interaction are biologically significant.

*Evidence in IEDB<sup>1</sup>:* We found experimental evidence for the significance for the virus motifs in the Immune Epitope<sup>2</sup> database (IEDB). The pattern IVGG from the *Hepatitis-C* motif in Fig. 5 is found in 53 epitopes. From the *Ebola* motifs in Fig 3, we find that TLAT is part of six different epitopes, SLTT appears in three epitopes. PLIK, SLLL from the *Influenza* motif are also found in many epitopes. Finding that the virus k-mer patterns predicted by our method are recognized by human antibodies is a further validation of its performance. Further, using higher dimensional  $k$ -mers (where  $k=7, 8, 9$ ) as features in our model will give motifs from which complete epitopes can be derived. Our model thus has applications in epitope prediction as well, where conventional methods consist of scanning all possible  $k$ -mers from protein sequences to identify likely epitopes.

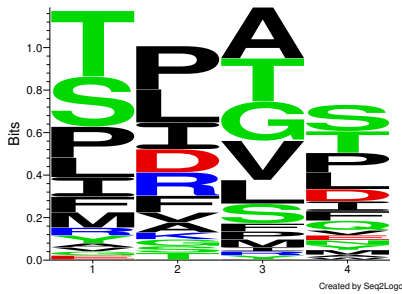


Figure 3: Sequence motif from top four-mer features specific to *Ebola* proteins

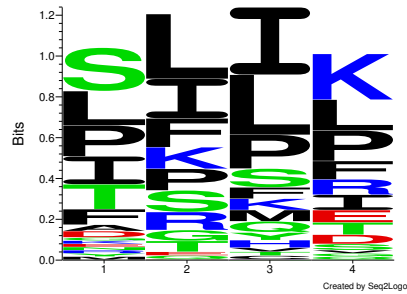


Figure 4: Sequence motifs specific to *Influenza* proteins that are also important to interactions.

<sup>1</sup>[www.iedb.org](http://www.iedb.org)

<sup>2</sup>An epitope is a very short sequence from the virus that binds to human antibodies

## 4 Tri-mers

Below, we show the sequence motifs from the tri-mers found to be highly relevant to predicting interactions between human and viral proteins.

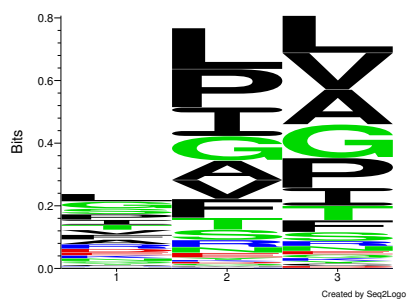


Figure 5: Sequence motifs specific to *Hepatitis-C* proteins that are also important to interactions.

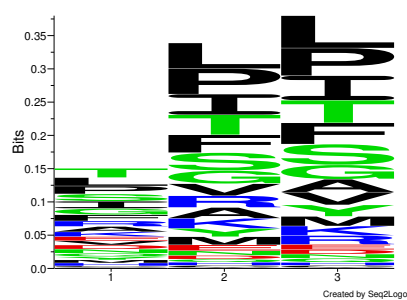


Figure 6: Sequence motif constructed from the top tri-mer features of human proteins